

MULTIMEDIA



UNIVERSITY

STUDENT ID NO

--	--	--	--	--	--	--	--	--	--

MULTIMEDIA UNIVERSITY

FINAL EXAMINATION

TRIMESTER 2, 2017/2018 SESSION

TDS 3301 – DATA MINING

(All Sections / Groups)

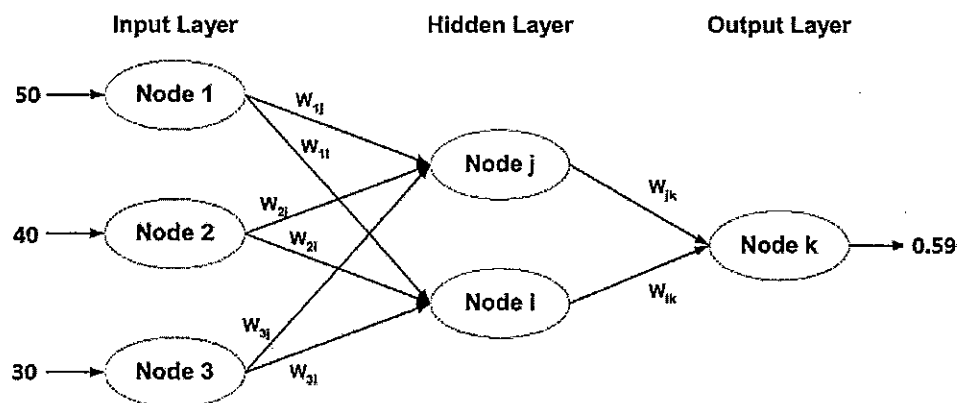
05 MARCH 2018
2:30 p.m – 4:30 p.m
(2 Hours)

INSTRUCTIONS TO STUDENTS

1. This Question paper consists of 4 printed pages including cover page with 4 questions only.
2. Attempt **ALL** questions. All questions carry equal marks and the distribution of marks for each question is given.
3. Please write all your answers in the Answer Booklet provided

Question 1

- Provide an example of data mining application and describe in brief the data mining approach used in
 - market analysis (2 marks)
 - fraud detection. (2 marks)
- Are all patterns found using data mining approaches interesting? What are the criteria of interesting patterns? Use association rule mining to explain your answer. (4 marks)
- Sometime it is too heavy for human to look at all the possible patterns in a large data. Suggest a method to find only the interesting patterns. (2 marks)

Question 2

A Neural Network (NN) is given as above. The initial weights are $W_{1j}=0.2$, $W_{1i}=0.1$, $W_{2j} 0.3$, $W_{2i} = -0.1$, $W_{3j}= -0.1$, $W_{3i}=0.2$, $W_{jk}=0.1$ and $W_{ik}=0.5$. Question (a) to (c) are about feed forwarding of this NN. Give your answer for the following questions up to 3 decimals only.

- Normalise the inputs to the nodes using min-max normalization, where the minimum and maximum values are, 10 and 50, respectively. Why normalization is important in training a NN? (3+1 marks)
- The inputs are mapped with the multiplication of vectors and a sigmoid function. Assuming output from node j and node i are, 0.593 and 0.531, respectively. What is the output from node k ? (2 marks)
- Compute the observed error at the output layer of the NN. (1 mark)
- What is the best rule to the best number of hidden layer nodes? (1 mark)
- Suggest a solution if a trained NN has an unacceptable accuracy. (2 marks)

Continued...

Question 3

The data for an attribute X are: 3.13, 4.53, 4.98, 5.09, 5.44, 6.11, 6.50, 6.68 and 7.22.

- a) Draw a boxplot to show the distribution of the data. Label the five important values of the data in the boxplot. (6 marks)
- b) Are the data normally distributed? Plot a Q-Q plot and check using the plot by pairing x_i to f_i , where $f_i = (i - 0.5)/N$. (4 marks)

Question 4

	X	$\neg X$
Y	100	1,000
$\neg Y$	1,000	100,000

- a) The 2-way contingency table of two items, X and Y , is as shown above.
 - i) An association rule is generated, $X \rightarrow Y (s, c, l)$. Calculate the support (s), confidence (c) and lift (l). (3 marks)
 - ii) What is the function of lift measure in association rule mining? (1 mark)
 - iii) Is the lift measure suitable in this case? Why or why not? (2 marks)

a	b	←Classified as
7	2	a=yes
4	1	b=no

- b) Fill in the following table based on information in the confusion matrix above.
 - i) How many records are correctly classified? (1 mark)
 - ii) The Recall is the proportion of examples which were classified as class x , among all examples which truly have class x . So, what is the Recall for class *no*? (1 mark)
 - iii) The Precision is the proportion of the examples which truly have class x among all those which were classified as class x . What is the Precision of class *yes*? (1 mark)
 - iv) Usually what will happen when you try to improve either Recall or Precision? (1 mark)

Continued...

Formulae:

Min max normalization, $new\ value = \frac{original\ value - minimum\ value}{maximum\ value - minimum\ value}$

Sigmoid function, $(x) = \frac{1}{1+e^{-x}}$, where e is 2.718282.

Back propagation of errors at output layer, $error(k) = (T - O_k)O_k(1 - O_k)$, where T is target output and O is the output of a node.

Support, s , is the probability that a transaction contains $X \cup Y$

Confidence, $c = \frac{sup(X \cup Y)}{sup(X)}$

Lift, $l = \frac{sup(X \cup Y)}{sup(X)sup(Y)}$